SOFTWARE MAINTENANCE: RESEARCH AND PRACTICE
J. Softw. Maint: Res. Pract. **10**, 279–303 (1998)

Research

Documentu: A Flexible Architecture for Documentation Production Based on a Reverse-engineering Strategy

CHRISTIANO de OLIVEIRA BRAGA*, ARNDT von STAA and JULIO CESAR SAMPAIO do PRADO LEITE

Departamento de Informática da Pontifícia Universidade Católica do Rio de Janeiro, R. Marquês de S. Vicente 225, Rio de Janeiro, RJ 22453–900, Brazil

SUMMARY

Good documentation is essential to the production and evolution of quality software. Based on a survey on program documentation, we propose a documentation architecture that aims to fulfil several requirements of an ideal solution. The architecture integrates a powerful transformation system (Draco-PUC) and a versatile CASE tool (Talisman) in order to produce documents using reverse-engineering strategies. We have implemented such an approach in a prototype tool called *Documentu,* which uses a database to store system-wide information and uses a standard HTML hypertext browser to help the access and navigation of system documentation. The prototype was used on a large scientific system that belongs to an oil company, with positive results.

KEY WORDS: documentation; transformation systems; CASE tools: software engineering; tool integration; reverse engineering

1. INTRODUCTION

One of the goals of a software process is to provide the necessary information to maintain evolving software systems. Usually, most of the system information is solely embedded in the current source code. However, other sources of information should be available in order to provide a better understanding of the software and its environment (Biggerstaff, Mitbarden and Webster, 1994; Selfridge, Waters and Chikofsky, 1993; Leite and Cerqueira, 1995). Good documentation, easily accessible and easy to read has been sought for a long time (Horowitz, Kemper and Narasimham, 1985; Garg and Scacchi, 1990). It is essential to software production, which employs several software engineers, working in a co-operative fashion.

CENPES is the research institute of Petrobrás, the Brazilian Government Oil Company.

* Correspondence to: Christiano de Oliveira Braga, Departamento de Informática da Pontifícia, Universidade Católica do Rio de Janeiro, R. Marquês de S. Vicente 225, Rio de Janeiro, RJ 22453–900, Brazil. Email: cbraga@inf.puc-rio.br

CCC 1040-550X/98/040279–25$17.50

Received 17 December 1997
Revised 11 May 1998

This centre uses several software systems developed both internally and by third parties. One of its projects, DOC, was created in order to improve the quality of the software products developed at CENPES, by making the maintenance process easier and increasing the quality of the documentation assets. The project is a partnership between LES—Software Engineering Laboratory at PUC-Rio and CENPES.

One of the deliverables of the DOC project was a set of guidelines for development and documentation of software code (Staa *et al.*, 1995, p. 5). However, practical experience has shown us that the guidelines would only produce the expected feedback if they could be assisted by a software tool.

We have surveyed some of the existing proposals for documentation tools at the code level, and based on their strong and weak points, we propose our architecture for a tool that implements automatic support to the guidelines. Our proposal uses HTML (hypertext mark-up language) and relates the information in terms of the whole system, not only for individual programs. The tool generates documentation for annotated and unmarked code as well.

We have implemented our proposal in a tool called *Documentu* (Braga, Staa and Leite, 1997, pp. 9–13). The main objectives of the tool are:

- help software personnel in the maintenance task;
- aid the software users in realizing the features and facilities of the product if the software being developed is in a library; and
- help a new developer to understand the software system with which he is going to work.

Documentu ideally uses code tag-marked as defined by Staa's guidelines, like naming variables or functions and inserting annotations in comments to document design decisions. The guidelines also provide for object-oriented programming and testing. The guidelines deal with higher abstraction levels as well, making suggestions on how to structure software projects. However, *Documentu* can also produce documentation from bare source code. This is possible because we generate software documentation based on both the source-code grammar and the mark-up grammar. We have used the transformation system of Draco-PUC (Leite, Sant'Anna and Freitas, 1994) to recover system information from code and mark-ups. The Talisman CASE tool (Staa, 1993, p. 1.1) was used to store, validate and format the information. Finally we use a standard HTML browser to access the resulting documentation.

The transformation system plays the role of the extractor, analysing and retrieving code information. This information is extracted using syntax analysis and, if needs be, mark-ups. This information is passed to the CASE tool that stores it in a repository. This repository has database management capabilities and is able to provide a series of consistency checks that can be applied to the entire system. Using the information in the repository, the CASE tool generates an HTML file by means of special procedures called linearizers.

Before detailing our proposal we make a brief survey of existing documentation proposals (Section 2). In Section 3, we describe the tool architecture. In Section 4, the *Documentu* tool is presented. Section 5 reports on the results of applying *Documentu* to

working scientific software. Section 6 presents what can be generated with the current version of the tool. We conclude the paper in Section 7, referring to related work and reporting on our plans for the evolution of *Documentu* and its use.

2. A BRIEF SURVEY ON DOCUMENTATION PREPARATION

2.1. Document standards

There is no doubt of the importance of presenting software artefacts as structured and readable documents. Nowadays, this issue is confronted with a plethora of new possibilities through the widespread diffusion of the World Wide Web and its *lingua franca*, HTML.

Our work was based on this perception and on key literature, which we classified under the following categories: standards, literate programming, using mark-up languages and databases, and recent documentation tools.

SGML (Goldfarb, 1990, pp. 217–494) is the ISO-8879 ISO standard for structured documents. The standard describes a language for document description. This language defines tags to be embedded in the text, in order to structure the document. The tags have the form <cmd> to begin a structure and <\cmd> to end it. A simple example would be:

```
<Paragraph>
This an example of a paragraph tag in SGML
< \Paragraph>
```

A style sheet can be applied over the document to produce a pretty-printed view of the document, following a document DTD (Document Type Definition). A DTD defines a structure for a class of documents. Figure 1 shows an example of DTDs.

A document created following the SGML DTD defined in Figure 1 would be composed of MANINFO (managerial information), USEINFO (use information) and IMPLEMENTINFO (implementation information). In the example of Figure 1 we have detailed the managerial information that could be composed of Project, FileType, IdNumber, Version, DataAp, and Author. Among those, IdNumber is optional and Author may include it one or more times. All documents created using this structure would necessarily have those sections.

```
<! -- ELEMENT CONTENTS -- >
<!ELEMENT DOCUMENT( MANINFO, USEINFO, IMPLEMENTINFO)>
<! -- ELEMENT MANINFO -- >
<!ELEMENT MANINFO ( Project, FileType, IdNumber?, Version, DateAp, Author+)>
<!ELEMENT Project ( #PCDATA )>
<!ELEMENT FileType( #PCDATA )>
<!ELEMENT IdNumber( #PCDATA )>
 ...
```

Figure 1. Example of an SGML DTD

HTML (World Wide Web Consortium, 1995) is based upon a subset and application of the SGML standard. The language is used to create portable hypertext documents. The HTML mark-ups can represent hypertext news, electronic mail, documentation, hypermedia, options menu, databases query results and structured documents with graphics. HTML has been used in the World Wide Web since 1990. The most important difference between HTML and SGML is that the former has the capacity to refer to other documents through links defined with mark-ups. Another relevant difference between SGML and HTML is that the programmer cannot develop different DTDs in HTML. All HTML documents have to follow a fixed structure, defined by the HTML standard.

The Office Document Architecture (ODA) (Appelt and Tetteh-Lartey, 1993; Brown, 1989) is an object-oriented framework, created to be a standard for document exchange. ODA, like SGML, distinguishes between the logical structure and the layout structure of the document (see Figure 2).

While SGML and ODA have different approaches to solve the same problem—documentation exchange—both of them emphasize a very important issue in documentation: the essential difference between the document structure and the layout structure. Also, both of them fail in supporting the creation of hypertext documents.

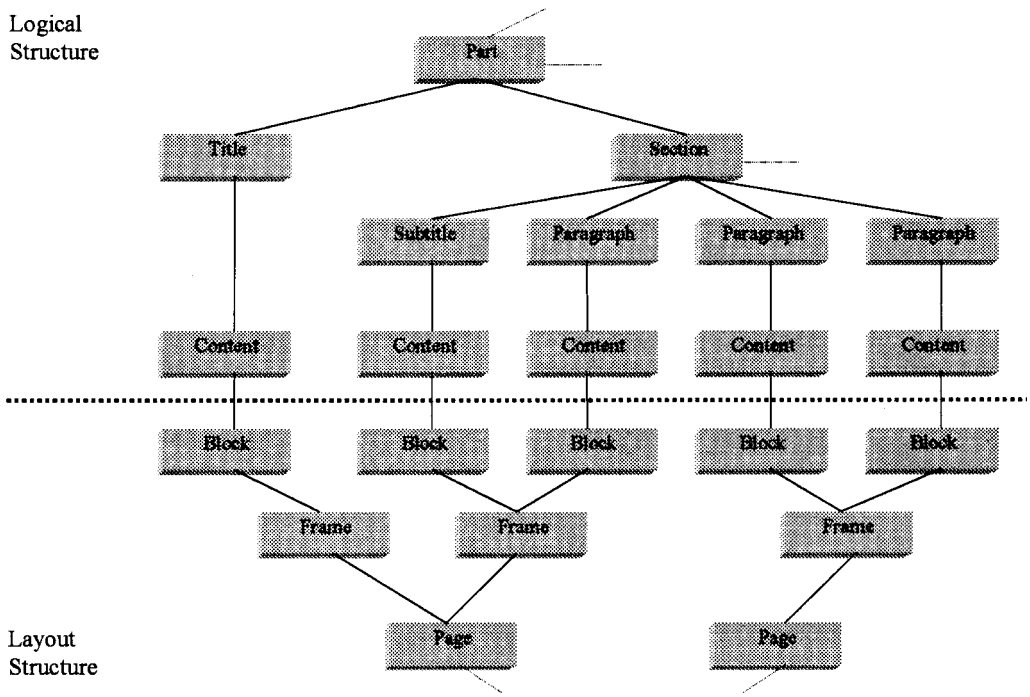

Figure 2. Diagram of the ODA architecture

2.2. Literate programming

Knuth (1984) proposed the joining of a document format language to a programming language in an approach called Literate Programming. In WEB—Knuth's implementation of the Literate Programming proposal—he uses the well-known languages T_EX and Pascal. To develop a system using Literate Programming, the programmer must, at first, write a WEB file, which is the document that describes the system being developed. This file is composed of the program description with chunks of code. The next step is to apply a processor called WEAVE to produce a T_EX file.

The use of the Tangle processor sets up the Pascal code files to be compiled. In this way documents tend to be consistent, since they come from the same source. Figure 3 shows the WEB development process. Knuth's proposal is indeed very interesting, since the main idea is to develop readable code for other human beings and not just code to be compiled. But the main problem is that Literate Programming was conceived to describe algorithms and not to produce system-wide documentation. The consequence is that the final documentation remains related to just a single module. The cross-references are bounded to the document edge, and they have to be explicitly created by the developer. Another problem is that T_EX does not provide support to document structure definition, like SGML DTDs, in such a way that format mark-ups are embedded in the document together with the text.

2.3. Mark-up languages and databases

Cowan presents a more elaborated solution (Cowan *et al*., 1994, p. 2) than the one proposed by Knuth. In Cowan's approach, SGML mark-up language is used to embed specification information in the code. A database is then filled with code information and, through the use of database resources, this information can be viewed in hypertext format. All documents resulting from the development process could be in SGML format linked to each other inside a textual database (Cowan *et al*., 1994, p. 5). All the information could be accessed through different views in a WYSIWYG (what you see is what you get) environment, allowing database update through those views.

Cowan *et al*. (1994, pp. 8–10) brings to light some important issues concerning the preparation of software systems documentation:

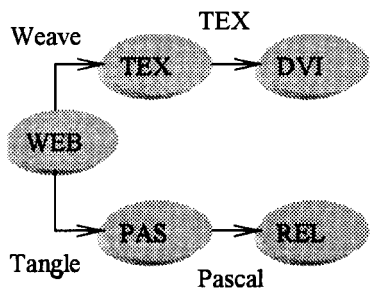

Figure 3. Generating WEB documents with TEX and Pascal

- the need for a document logical structure, defined by a standard in order to achieve document portability;
- the need for databases, to store the system-wide code information, allowing validations over the gathered data;
- the use of a hypertext browser to navigate across the information.

2.4. Documentation tools

There are several tools available to support the generation of software development process documentation, but most of them only produce documentation automatically without checking for consistency with the code.

Java (Gosling, Joy and Steele, 1996, p. 1) is a programming language developed at Sun Microsystems to support the development of portable and reusable software components. Figure 4 shows the browsing of a document generated by JavaDoc (Flanagan, 1996, pp. 220–221), a tool that generates HTML documents from tag-annotated Java source code. The document in Figure 4 was generated from the annotated code presented in Figure 5.

JavaDoc formats public and protected information from classes, interfaces, constructors, methods and fields. Figure 5 shows an example of JavaDoc tags. HTML tags can also be used to format comments. User-defined cross-reference tags, such as `@see` (not shown

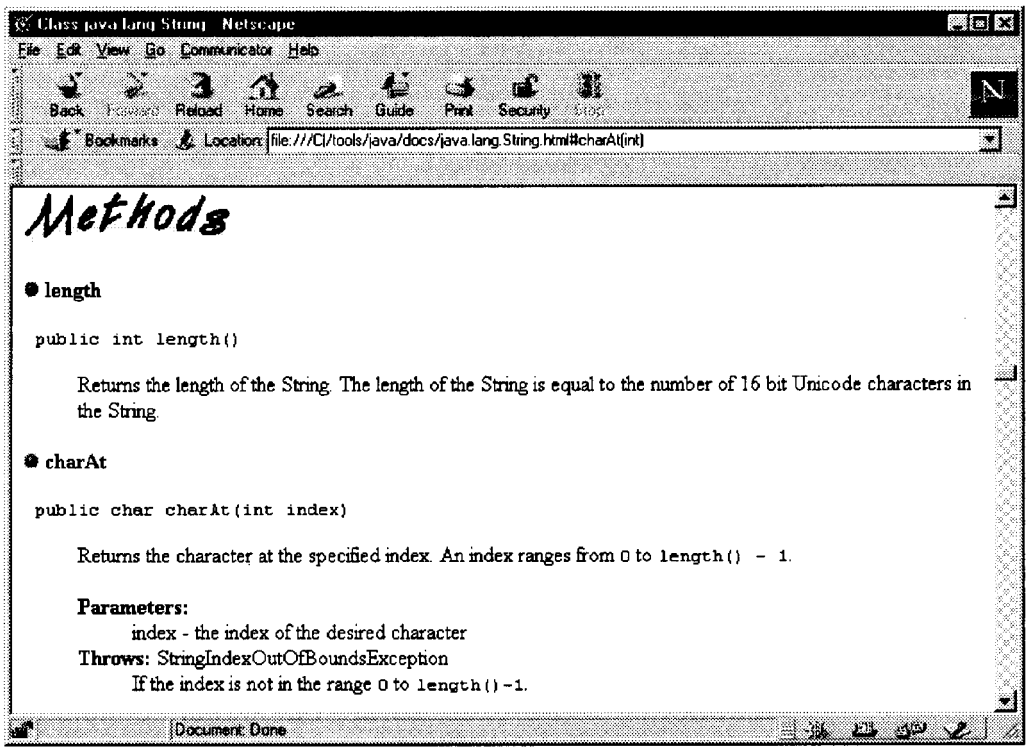

Figure 4. Example of JavaDoc document browsing


```
/**
* Returns the length of the String.
* The length of the String is equal to the number of 16 bit
* Unicode characters in the String.
*/
public int length()
{
    return count;
}
/**
* Returns the character at the specified index. An index ranges
* from <tt>0</tt> to <tt>length() - 1</tt>.
* @param index the index of the desired character
* @exception StringIndexOutOfBoundsException If the index is not
*   in the range <tt>0</tt> to <tt>length()-1</tt>.
*/
public char charAt(int index)
{
    if ((index < 0) || (index >= count))
    {
      throw new StringIndexOutOfBoundsException(index);
    }
    return value[index + offset];
}
```

Figure 5. Example of annotated Java code

in the example), may be manually inserted. JavaDoc creates documentation (an index, a class hierarchy and description files) for Java source files or for packages.

Microsoft provides an unsupported tool called AutoDuck. This tool works much like JavaDoc. It uses mark-up tags in the source code to generate documentation. This documentation is generated in RTF (Rich Text Format—a Microsoft standard available for Windows, OS/2 and MacOS platforms). RTF is the format used by the Microsoft Help Compiler to generate Windows Help files. The AutoDuck tag @class (see example in Figure 6) defines the name of a class and @cmember defines a class member. Since the class in the example is a template, `@targs` defines the class templates arguments. Figure 7 shows the browsing, using the Windows help system, for the example presented.

AutoDuck provides the capability of parsing user-defined tags. JavaDoc parses only its predefined tags. Both of them are easily integrated into the building process through `makefile` invocation, improving documentation consistency with the source code. If the mark-ups were consistently updated and the documentation generated whenever the code changed, then the documentation would always be consistent with its associated source code.

AutoDuck does not support system-wide documentation. It is only concerned about local modules and cross-references have to be created manually by the developer, generating possible sources of inconsistency. Neither AutoDuck nor JavaDoc perform consistency checking in the retrieved information.


```
/* @doc
 * @module TEMPLATE.CPP - Some Template Expls.
 * This module shows how to code Autoduck tags
 * template classes and member functions.*/
//@class Template class
//@tcarg class | T | A class to store in stack
//@tcarg int | i | Initial s of stack
template<class T, int i> class MyStack
{
    //@cmember The top
    T*  pStack;
    //@cmember The stack
    T StackBuffer[i];
    //@cmember The count
    int cItems = i * sizeof(T);
  public:
    //@cmember Constructor
    MyStack( void );
    //@cmember Add one
    void push( const T item );
    //@cmember Remove one
    T& pop( void );
};
```

Figure 6. Example of AutoDuck tags

2.5. Some important issues about documentation preparation

The main concept to keep in mind is that software programs have to be written for human beings, and not only to be processed by computers (Knuth, 1984). The existence of a document logical structure, defined using a standard in order to achieve document portability is a major issue (Goldfarb, 1990, pp. 402–433). To store system-wide code information and thus allow checking among information fragments, requires the use of a database (Cowan *et al*., 1994, p. 8). Another important issue is the use of a hypertext browser to navigate through the information (Bigelow, 1988; Rajlich *et al*., 1990). The use of hypertext provides an easy way to access on-line information, as well as allowing information layering, such that information can be structured.

3. THE PROPOSED ARCHITECTURE

3.1. Draco-PUC

Our proposal is geared to marked code, but it also treats unmarked code. This makes possible the generation of the documentation based solely on a system's source code (currently C, C++ and FORTRAN 77). It is important to emphasize that the tags only complement the information already captured by the tool. *Documentu* is a tool that instantiates our proposal. The marks treated by our proposal are based on coding guidelines. Section 4 describes these coding standards. If there is no documentation available in the

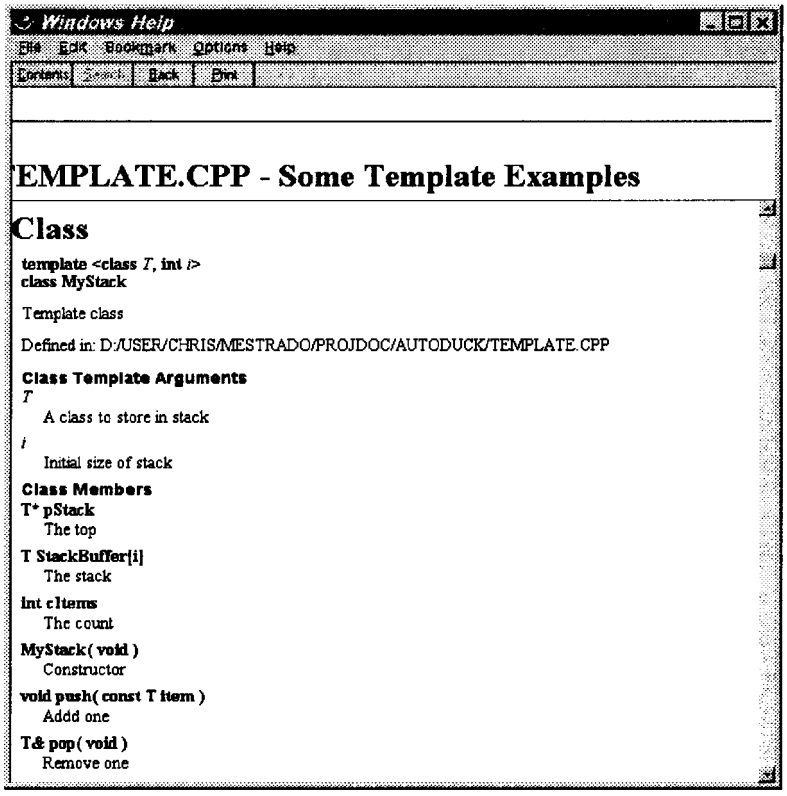

Figure 7. Example of AutoDuck browsing

code, the information shown in the generated documentation is what can be inferred from the source-code language syntax through parsing and transformation.

The information extractor (see Figure 8) parses the source code and produces information that is used by the information formatter system. The formatter system then produces a hypertext system in HTML. The first part of the process, the information extractor system, uses the Draco-PUC (Leite, Sant'Anna and Freitas, 1994) transformation capabilities. The

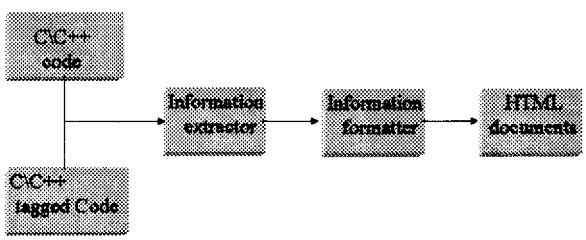

Figure 8. Outline of the proposed architecture

second part of the architecture, the information formatter system, is based on a repository-based CASE tool called Talisman (Staa, 1993, p. 1.7).

Draco-PUC (Leite, Sant'anna and Freitas, 1994) is a transformation system that implements the Draco paradigm (Neighbors, 1984). This paradigm performs software development by creating and reusing domains, at a high level of abstraction. An essential part of a domain description in Draco-PUC is a grammar definition written in a BNF-like style. Using this grammar as input, the **pargen** subsystem (**par**ser **gen**erator), produces the domain parser. The unparser or pretty-printer is also created when the grammar is annotated with layout directives.

A special abstract syntax tree (DAST—Draco Abstract Syntax Tree) is created to store the information retrieved by the parser. Once this structure is created, it is possible to apply the unparser to view the information in the DAST as described in the layout directives. It is also possible to apply transformations to map the information in the DAST to a new structure over the same domain (intra-domain transformations) or to a new domain using inter-domain transformations.

A Draco-PUC transformation has a left-hand side (LHS), which defines a searching pattern and a right-hand side (RHS), which defines a replacement pattern. Control points and transformation properties are also available. These kinds of features are responsible for driving the transformation engine in a bottom–up search-and-replace way. The control points of a transformation are:

- **Pre-match**—executed when the engine tries to execute the transformation.
- **Match constraint**—executed every time a pattern variable is bound against a program fragment when the transformation matching is being attempted.
- **Match failure**—executed when the transformation matching fails.
- **Post-match**—executed right after the successful matching of the transformation.
- **Pre-apply**—executed before the replacement of the LHS by the RHS.
- **Post-apply**—executed after a pattern is substituted.

Transformation rules are grouped in transformation sets, which also have control points (initialization and end) and properties. Sets of transformations are encapsulated on transformers. Transformers have three control points: declaration (global object declaration), initialization and end.

Draco-PUC transformations can be in-place transformations, when the substitution pattern is replaced in the source, or they can work with workspaces. Transformations that use workspaces are non-destructive and the replacement patterns are not instantiated in analysed source, but rather in an external workspace. When using workspaces, it is also necessary to define templates, which are Draco-PUC transformation language structures that allow the developer to instantiate the RHS in target workspaces.

The use of workspaces can be viewed as an asynchronous replacement, where the RHS is written in a temporary space for further use. As an example of such use, consider the creation of a relation between one element that was already found by the analyser and another that is to be found. A workspace should be used to create this late association. An example will be presented in Section 4.1.

3.2. Talisman CASE tool

Talisman (Staa, 1993, p. 1.7) is a repository-based CASE tool. It provides an environment including predefined database schemas, called *Talisman Definition Languages* (Staa, 1993, p. 1.11) that give the semantics to the existing information inside a software base. Talisman supports languages for structured analysis, and diagrams like the data-flow diagram for data modelling, and the entity-relationship diagram for information systems modelling. Talisman also allows the definition of such languages, thus making customization an easy task.

A software base (Staa, 1993, p. 1.7) is a database composed of textual fragments and references to textual fragments inside the database. A definition language (Staa, 1993, p. 14.1) is characterized by the relations among its elements (modules, classes and functions, for example) and also *form programs* (validators and linearizers). Those *form programs* could be loaded at run time to customize the environment. The validators check for consistency of the data inside the software base, whereas the linearizers produce reports with the information contained inside the software base. When a software system is developed, editor forms are used to populate the software base with specification text fragments. Each text fragment can have different roles depending on which definition language is being used. For instance, a data repository in a data-flow diagram can be represented as a table in an entity-relationship diagram.

Once the software base is populated, the validators have to be applied to perform consistency checking with the information inside the base. With the information validated, the linearizers are applied to produce documentation or code.

3.3. Integration of the tools

The information extractor system (Figure 8) is set up on Draco-PUC by means of two domains: the C/C++ domain and the Talisman domain. The C/C++ domain was already available as a Draco-PUC executable domain. This domain has been used by several people in different projects, indicating its robustness. However, it was necessary to extend the domain in order to make possible the analysis of special tags.

Talisman has an importing feature that allows the CASE environment to populate an existing base with information provided by imported files. The rules that guide the creation of this file can be written as a grammar. Using this grammar, the Talisman domain was created at the Draco-PUC machine.

The process starts with the analysis of the source code. Once the code of a given module is analysed and the Draco-PUC abstract syntax tree is created, a set of transformations, created specially to implement this approach, are applied over that internal format in order to create the Talisman import file for that module. This operation is repeated for each module described in a Draco-PUC script file (DSF file). A Draco script file is written for the whole system, capturing all the modules it comprises.

Figure 9 shows this process. In the first phase, tagged (or not tagged) code is analysed by the transformation system, generating the import files to the Talisman CASE tool. The second part of the process is restricted to Talisman. Each import file outputted by the Draco-PUC machine is then imported to the Talisman environment. Once all the files are

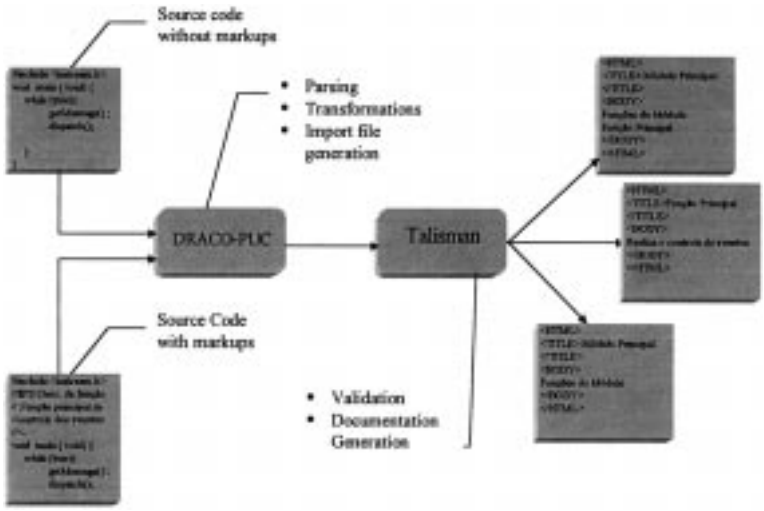

Figure 9. Diagram of the Documentu *execution process*

imported, the validators are executed, checking the data consistency inside the base. They look for text fragment completion and cross-reference validation. A log is created with the inconsistencies that have been found. When the base is validated, the linearizers are executed to create the HTML documents.

4. DOCUMENTU

4.1. Role of Draco-PUC in the architecture

We describe *Documentu* by focusing on details of its architecture, its execution and the results produced. The tool is based on three Draco-PUC domains, C++, FORTRAN and the Talisman domain, and on the repository defined in Talisman. Below we give details on how these parts are organized.

As cited before, the basis for a Draco-PUC domain description is its language grammar description. A domain-language grammar description is written in an extended-BNF style. An example of such grammar descriptions is shown in Figure 10, which presents the CPP grammar format for a C++ language fragment responsible for the description of class heads like:

```
class Cwindow {. . .}
//Or . . .
class CMyWindow : CWindow {. . .}
//Or . . .
class CAnotherWindow : CWindow, CBaseWindow {. . .}
```

Just like with C++, the Talisman import file grammar description was also encapsulated


```
class_head : class_key class_name? base_spec?
        ;
class_key  : 'class'
        | 'struct'
        | 'enum'
        ;
base_spec    : ':' base_specifier+','
        ;
base_specifier: class_name
        | 'virtual' access_specifier? class_name
        | access_specifier 'virtual'? class_name
        ;
class_name   : CLASS_NAME
        ;
access_specifier : 'private'
        | 'protected'
        | 'public'
        ;

import_file  : import_entry*
        ;
import_entry : command
        | comment
        | exc_obj
        | exc_attrib
        ;

command : 'DA' category name attribute
        ;
category     : STRING
        ;
name         : STRING
        ;
attribute    : 'Name'
        | text_field
        | alias
        | relation
        ;
text_field   : 'Text' ( ID | NUMBER ) lines_of_text;
```

Figure 10. Fragments of the CPP and Talis grammars

in Draco-PUC. An excerpt of this grammar (Talis), which describes textual import entry for the repository is also shown in Figure 10. An example of a textual import entry is the one *Documentu* uses when defining descriptive long names for classes in the Talisman repository.

After the grammars were defined, transformations were created to retrieve the information from the C/C++ code into the Talisman domain in such a way that it could be used in the CASE tool. The retrieved information comes from the language (C/C++) analysis, and also from the embedded tags following the specification information, defined in Staa *et al.* 1995, pp. 11–32) as part of a set of C/C++ coding standards. The embedded tags


```
////////////////////////////
//$CNAME Class long name
//VMS - Segment
//$CD Class description
//Segments are storage units. Each Segment is related to a
//file. Segments are composed of pages of fixed size. Segments
//can be created, deleted, opened or closed just like files.
//////////////////////////////
class VM_Segment
{
     ...
};
```

Figure 11. Examples of embedded tags

appear as comments, so they do not change the compilation process. Figure 11 presents an example of such tags embedded in C++ code.

In *Documentu*, there are two types of transformations: tag matching and structure matching (concerning language structures). The tag-matching transformations retrieve the design information embedded in the code as comments (Staa *et al.*, 1995, pp. 11–32). The structure-matching transformations are intended to retrieve information about the program structure using the code itself as the main source of information. The documentation structure is created using the information retrieved by these transformations.

An example of a tag-matching transformation is shown in Figure 12. The transformation looks for a class name tag specification (`$CNAME`) and then stores the proper information in a workspace (`WSClassLongName`). Following Staa's commenting standard, each class should have a long name in order to provide a better understanding of the abstraction. Later in the transformation process, a transformation that looks for classes, as in Figure 13, will be executed and will use the name stored in that workspace to set the long name attribute for the associated class. The class that will have the attribute set is the next class following the comment describing the `CNAME` attribute.

```
Transform GetClassLongName
  Lhs: {{dast cpp.tag
       //$CNAME [[IDENTIFIER* tag_specifiers]]
       [[TAG_LINE long_name]]
  }}
  Post-Match: {{dast cpp.statement_list
       TEMPLATE("StoreCurrentClassLongName");
            TRANSPORT_VALUE("long_name");
            PLACE_AT("WSClassLongName");
       END_TEMPLATE;
  }}
  Template StoreCurrentClassLongName
  Rhs: {{dast talisman.lines_of_text
       [[LINE long_name]]
  }}
```

Figure 12. Example of a tag-matching transformation


```
Transform Class
  Lhs: {{dast cpp.class_head
  class [[CLASS_NAME short_name]]
  }}
  Post-Match: {{dast cpp.statement_list
       TEMPLATE("ClassLongNameDeclaration");
             TRANSPORT_VALUE("short_name");
             MOVE("WSClassLongName","class_long_name") ;
             PLACE_AT("WSTalismanImportFile");
       END_TEMPLATE;
  }}
  Template ClassLongnameDeclaration
  Rhs: {{dast talisman.import_entry
       DA "Classes" [[ID short_name]] Text 30
       [LINE class_long_name]]
  }}
```

Figure 13. Example of a structure-matching transformation

The Class transform is an example of a structure-matching transformation. This transformation is responsible for generating the text that will create a class element in the Talisman repository (this is like creating a record in a relation database). This transformation also gathers other information from several workspaces concerning the class being analysed, and generates the text for what is gathered. This text will be imported by Talisman to populate the repository. An example of the generated text would be:

DA “Class” “VM_Segment” Name

DA “Class” “VM_Segment” Text ClassName

VMS – Segment

#@#

The first line creates a Talisman class element named VM_Segment. The second line sets the `ClassLongname` attribute of `VM_Segment` to `VMS-Segment`.

In the current version of the tool, the set of transformations cannot be expanded by a regular user. Once the transformation is written and compiled, it is only necessary to add it to the Draco–PUC script that applies the transformations. The problem is to write the transformations. The user needs to understand the C++ grammar—written in a BNF-like syntax—to learn Draco-PUC transformation language and know the structure of the already implemented transformations in order to use the retrieved context information. Also, the compiler used to generate Draco-PUC has to be the same for all transformations.

4.2. Role of Talisman in the proposed architecture

A new definition language has been implemented in Talisman to support the *Documentu* project. It is called Program Documentation language. Figure 14 shows a diagram of the language schema. This new language is based on Modular Programming (Staa, 1993,

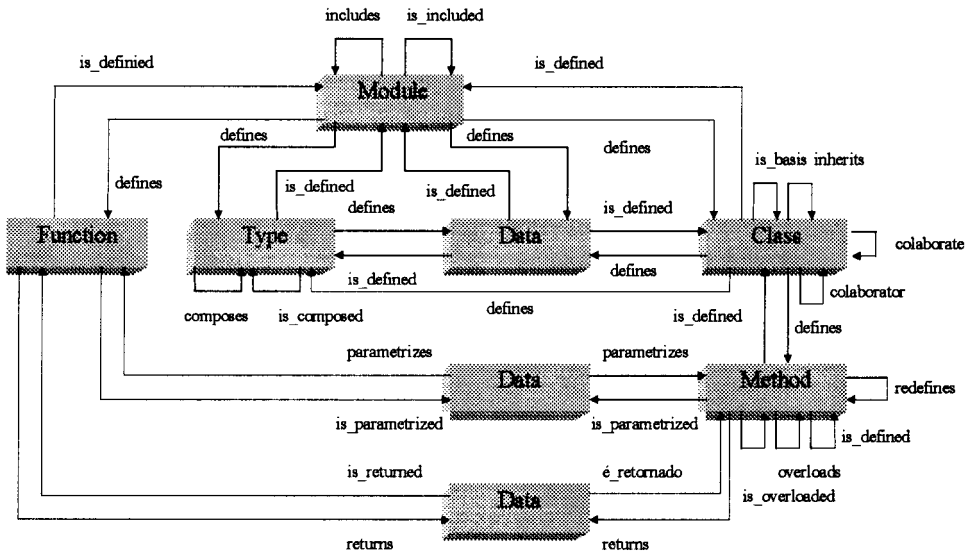

Figure 14. Model for Program Documentation definition language in Talisman

p. 22.1), but the information it manages is related to source-code documentation. When a file is imported, the information inside the file populates the software base.

Once the base is filled, the validators should be applied followed by the linearizers. In *Documentu* a linearizer is responsible for the creation of the hypertext system in HTML and just like the validator, is a *form program.* A *form program* is written using the language available in Talisman, to create programs, that manipulates the software base text fragments. The validators check the information inside the software base in order to verify its consistency. The linearizer navigates inside the software base looking for specific text fragments and relations and generates a file with the text fragment information in HTML format. An example of a Talisman form is shown in Figure 15.

```
Forall Classes do
    Frm "Create BeginSplittingSeq"(Convert Sequence [Name]);
    Frm "Create BeginHTMLFile";
    seqTitle = ConcatSeq ( "Class ",Convert Sequence [Name]);
    Frm "Create Title" (seqTitle,1) ;
    Frm "Create Reference" ("classes.htm","System Classes") ;
    if Exists ([Relation Module]) then
      Forall Module do
        seqClasseHead = ConcatSeq (Convert Sequence [Name],".htm");
        Frm "Create Reference"(seqClasseCabec,ConcatSeq("Module",
                                  Converte Sequence [ Name ]));
      end ; /* Forall */
   else
   end ; /* if */
end ; /* forall */
```

Figure 15. Talisman form excerpt

Figure 15 is an excerpt of the class linearizer. The first two lines of the code, after the `forall` statement, just call two other forms (like procedures, in this case). The first form program creates annotations that will be used by the file splitter at the post-processing phase. The second form program produces the mark-ups that all HTML files should have (<HTML>). A line with the class name is created next, using the HTML mark-up <H1>, since 1 was passed as a parameter to `"Create Title"` form. A reference to the file `classes.htm` is also created using the form `"Create Reference"` with `"System Classes"` as the text reference. The `if` clause checks whether a class has a relation with some module or not. If it does, a reference to the related module will be created writing the module name in the reference.

First the linearizer creates the documentation structure and the index and only then is the documentation created for each element of an entity. Viewing the documentation as a graph, first the raw nodes with the links are created, then the node attributes are filled.

One may be thinking why Talisman was used, since Draco-PUC could generate all HTML files itself. Talisman was used for the same reasons a database is used in software development. We wanted to be able to generate the documentation using a high-4GL-like language construction. Talisman *form language* possesses these constructions, as shown in Figure 15. The language provides constructions like `forall, exists` and also provides the capability to define new procedures. If Talisman were not used, we would have to manipulate text in files generated by Draco-PUC. Also, Talisman is shareware—a pre-requisite since we want to distribute the *Documentu* tool, well documented and with the people who built it available to implement the features we wanted, like creating a command line version of it.

Documentu uses Lua script files to drive all the execution processes. Lua is a portable script language developed at the Department of Informatics of PUC-Rio (Ierusalimschy, Figueiredo and Celes, 1996). The user only needs to call the main script file with the module name he or she wants to document. Once the main script file is called, the first thing it does is to call a pre-processing script file, responsible for calling the C++ pre-processor that expands the source code with the information inside the included header files. The pre-processed code is then analysed and further transformed into another file.

The file built by the transformation process is imported by Talisman, and is used to populate a software base with the information contained in the file. The HTML generator script file then calls Talisman to linearize the software base. The post-processing batch file calls the file splitter and the character converter to generate the multiple HTML files and to generate the HTML entities for the accentuated characters respectively. When all this is done, the user may invoke his or her HTML 3.0 (since the produced documentation is frame based) compliant browser to navigate through the documentation.

4.3. Hypertext

All the links are created using the extractor facilities of Draco-PUC (parser and transformations). The detailed information about each element (class and methods, for example) comes from the tags text embedded in the code. The information retrieved from the code analysis is used to create the hyperdocument structure: the index and hyperlinks cross-referencing system-entities, like modules, functions and classes. This structure is

what is created when no annotations are available in the code comments. The current implementation of *Documentu* generates a hypertext interface organized in two frames as shown in Figure 16.

In Figure 16, the frame in the left side is an index to system elements. The index is organized in tabs, like Windows selection tabs. The first one is the index to the system modules. Each bullet inside that frame represents a link to a module in the system. When a module is selected in the left frame, the big pane in the right changes, showing that module's information, like its associated classes, its functions, types and data. The same thing can be done with the other three tabs in the left pane, which represent the indexes to classes, functions and types of the system.

Using the links that appear inside the big frame, the user can navigate to the detailed information for a given system element. For example, consider the Shell module in Figure 16. The user can browse through information such as: classes in a module, functions, types and all the information defined in the model (such as illustrated in Figure 14) and

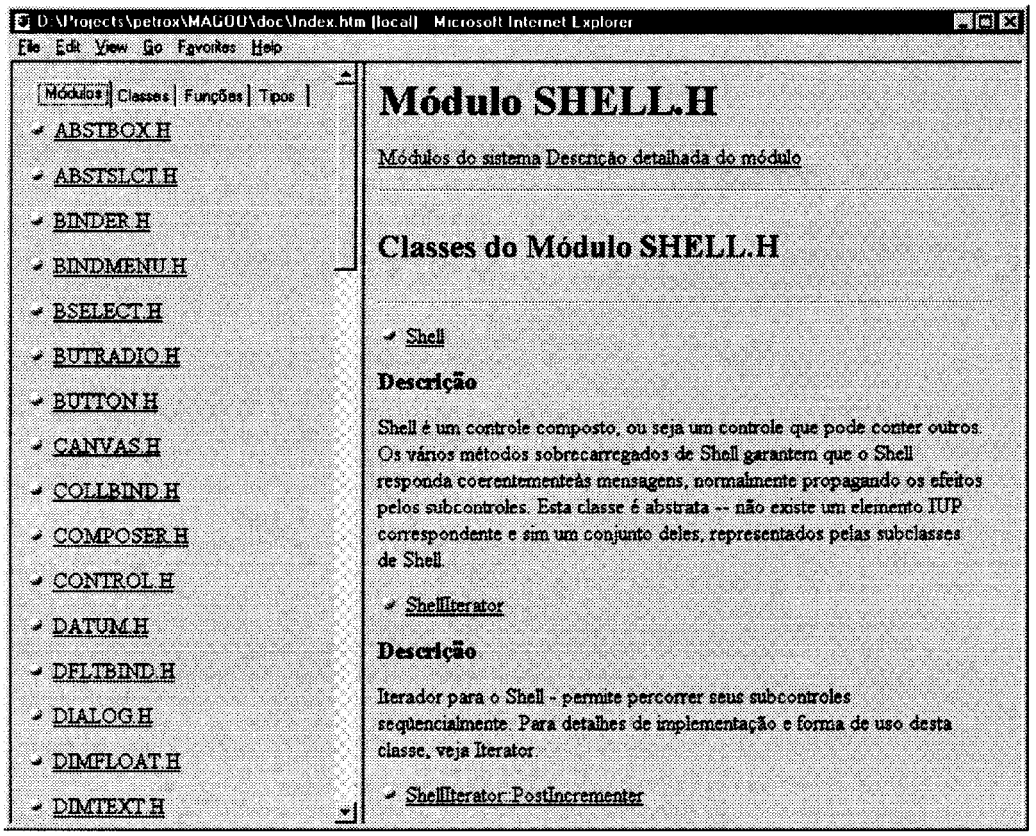

Figure 16. In this hypertext browsing, the generated text is written in Portuguese. 'Módulo Shell' means 'Shell Module' and 'Descrição' means 'Description'. There are two links below the big pane title: 'Módulos do sistema' meaning 'System Modules' and 'Descrição detalhada do módulo' meaning 'Detailed module description'

also, the information in the mark-ups, as defined in Staa *et al*. (1995, p. 11–32). Figure 16 presents a snapshot of the documentation browsing of the module `(Shell.h)` that has its classes annotated with mark-ups describing their functionality.

5. USING *DOCUMENTU* IN A REAL SYSTEM

The development of the *Documentu* tool started as part of the DOC project. This project, between LES (Software Engineering Laboratory at PUC-Rio) and CENPES (Brazilian Petroleum Company Research Centre), has the objective of increasing the quality of the software developed in that center.

In order to validate the tool, a large system being developed at CENPES was chosen to be the test case. The chosen system is called Petrox, a simulator for chemical processes. It is composed of an editor where the user may drop objects from a chemical process and simulate the execution of a particular instance of a process. The editor calls functions implemented in FORTRAN, which implement the algorithms needed to run the simulation.

Magoo GUI (Graphical User Interface Library) class library is a significant part of Petrox, currently with 72 header files in 5 057 595 bytes, in almost 208 552 lines of expanded code. We have been using Magoo as one of our test cases, since its code uses advanced features of the C++ language. Classlib—another relevant part of the Petrox project with 89 files and 369 kbytes—and Vix—a class library for visual object handling, with 43 files and 189 kbytes—were also used to test the tool.

Documentu took 30 minutes transforming the source code on a Pentium 100 running Windows NT 4.0 to produce the documentation, applying 17 609 transformations. The documentation was spread along 1 803 files using 1 128 953 bytes of disk space. These results sound pretty acceptable, considering the amount of information retrieved and that the documentation process is a batch process and may run in off-duty hours. Table 1 summarizes the results of running *Documentu* with Magoo as input.

Talisman linearizers have 32 functions and 2 012 lines of code. *Documentu* has 64 transformations implemented, distributed as follows:

- 43 concerning C++ classes;
- 30 concerning modules;
- 40 concerning functions;
- 61 concerning methods;
- 23 concerning structures, unions and enumerations.

Several features have been implemented in the tool since the delivery of the first prototype and many difficulties were encountered. The major problem found concerns the C++ parser. One of the first project decisions was that the symbol table generated by the currently available C++ parser—that ships with Draco-PUC—should be kept. The reason was because we wanted to evolve the project towards a reverse-engineering tool. This decision made a serious impact in the parsing phase, because the pre-processing of the source code became an obligation, since every type identifier was supposed to be in the symbol table.

Most of the parsing problems encountered were in system files included in the source

Table 1. Statistics obtained running *Documentu* **using the Magoo class library as input**

Module name	Size in kilobytes	#Tag Tfs.	#Lang. Tfs.	#Tfs p/ module	Module name	Size in kilobytes	#Tag Tfs.	#Lang. Tfs.	#Tfs p/ module
ABSTBOX.H	82 057	82	89	171	OBJBIND.H	94 502	75	79	154
ABSTSLCT.H	126 587	160	167	327	OBJSTRM.H	65 726	0	1 067	1 067
APPLIC.H	65 025	0	0	0	OPTBIND.H	18 322	126	143	269
BINDER.H	14 126	57	56	113	OPTBUT.H	83 856	165	168	333
BINDMENU.H	103 286	56	65	121	OPTMENU.H	103 870	57	66	123
BSELECT.H	87 889	180	181	361	PTRBIND.H	71 040	110	125	235
BUTRADIO.H	82 431	93	99	192	RADIO.H	84 653	96	104	200
BUTTON.H	78 854	229	226	455	SAFEREF.H	23 185	46	51	97
CANVAS.H	78 801	372	354	726	SELECTOR.H	132 599	218	225	443
COLLBIND.H	50 192	92	99	191	SEQBIND.H	72 085	122	137	259
COMPOSER.H	46 441	461	538	999	SHELL.H	79 648	218	207	425
CONTROL.H	71 931	0	708	708	SLCTDLG.H	122 726	66	73	139
DATUM.H	12 322	91	94	185	SLCTMENU.H	103 913	57	67	124
DFLTBIND.H	130 874	114	117	231	SRCBIND.H	31 041	127	141	268
DIALOG.H	93 217	309	299	608	STATUSLN.H	101 138	65	72	137
DIMFLOAT.H	63 637	196	194	390	STREAMBL.H	5 249	43	53	96
DIMTEXT.H	251 038	221	230	451	STRNG.H	37 297	246	255	501
DIMTYPES.H	4 573	0	0	0	STRVLD.H	70 635	94	101	195
ENTRBIND.H	16 555	85	94	161	SUBMENU.H	99 907	60	67	127
ENUMERAT.H	3 328	73	88	161	SYSDEPEN.H	8 062	0	0	0
FILL.H	73 488	47	54	101	TABLE.H	156 298	0	560	28
FLAGITEM.H	75 797	69	75	144	TASK.H	3 049	14	14	28
FLTVLD.H	51 331	193	187	380	TEXT.H	107 168	161	166	327
FRAME.H	81 415	85	89	174	TOGGLE.H	79 375	208	204	412
GERAL.H	1 362	8	8	16	TOOLBOX.H	85 174	59	69	128
HBOX.H	83 906	59	68	127	UNIT.H	53 308	345	378	723
HEAPVIEW.H	101 035	62	70	132	UNITBIND.H	129 936	90	93	183
INDEXDLG.H	110 243	60	69	129	UNITID.H	10 520	4	0	4
INTVLD.H	50 298	152	146	298	VALIDAT.H	40 816	208	180	388
LABEL.H	99 784	184	183	367	VBOX.H	83 966	59	68	127
MENU.H	98 398	89	91	180	XGETCH.H	3 473	2	2	4
MENUITEM.H	73 950	88	93	181	ZBOX.H	86 102	151	155	306
MFLOAT.H	14 446	91	94	185					
MINTEGER.H	14 216	91	94	185					
MULTLINE.H	104 036	181	182	363					
MULTSLCT.H	129 102	109	114	223					
NUMTYPES.H	121 888	0	23	23					
					Number of modules	69			
					Size of modules	5 057 595			
					Number of Applied transformation	17 609			

code by the pre-processor. The majority of our efforts were in making the parser as reliable as possible to support parsing included system files.

It was indeed a very difficult task, which is not fully completed, since the C++ language is not yet standardized. What should have been done—and will be in the near future—is to remove the symbol table and thus handle every type identifier simply as identifiers and not save context information. Any necessary context information will be handled by the transformation system. With this strategy, it will not be necessary to handle compiler-specific features of the language, sometimes implemented in system header files.

Another modification that could be done in the parser is to flatten some parts of the grammar to sequences of identifiers, making the C++ grammar more flexible in relation to its formal description. When using the whole grammar we have to be aware of language constructions that will not be handled by the transformation system, but that must be present in order to have the code parsed. This is another inheritance from the project decision of implementing a parser for a reverse-engineering tool.

Concerning the generated documentation, our experience has shown that using the language grammar to produce the structure of the documentation was pretty successful. Generating documentation from code that did not have any mark-up induced the user to annotate his or her code in order to produce a better documentation.

The system is presently installed in three different host systems at CENPES and it is still undergoing tests. The actual version is almost stable but still needs some technical support. We expect that it will reach desired stability in the next version provided the modifications discussed above are included.

6. RESULTS ACHIEVED

We realize that system documentation is a widely studied field with significant proposals and commercial products providing very powerful and complex solutions. We believe that our architecture contributes with the following:

- Flexible architecture. *Documentu* has an architecture that can be easily maintained and updated. Without changing the documentation generators, new programming languages can be supported by adding a new parser and transformers to the process. Also, the mark-up structure could be changed to follow different documentation guidelines, with different tag structures in the same way a new language is added. By loading new Talisman linearizers the documentation structure can be changed, so that, for example, instead of an HTML file, a postscript file could be generated.
- Compiler independence. Most of the currently available IDEs (Integrated Development Environments) provide features for code exploration. Our tool supports most of those features (like attribute, method, class and cross-referencing browsing) and it is compiler independent.
- System-wide documentation. The documented project can be as wide as the user wants, with references to as many modules and syntactic elements as the existing code has. Raw documentation, if no documentation tag is used at all, showing the system basic structure, can be produced.
- Documentation portability. This comes as a consequence of the use of a standard language.

It is also important to stress that by using the CASE tool Talisman (Staa, 1993, p. 14.1) it was possible to write form programs that could validate the integrity of the data relations (validators). After the analysis of the documentation, Talisman linearizers produce a hypertext document in HTML. The produced information will be extremely valuable to software personnel, mainly during maintenance. We understand that one of the strong points of our proposal is the low coupling among its components.

Also, there are some points we realize that must be reviewed in the next versions of the tool:

- Draco-PUC is still a prototype with some problems concerning usability and documentation. Creating the parser is not a very easy task. Talisman has an obsolete user interface. Creating a command line version, which also fits a lot better with our batch documentation process, has solved this. Also, the tool is only available in the Windows platform. We are working on a Unix version.
- The generated documentation is static. The user cannot 'ask questions' different from those defined in the form programs. We intend to add a feature in the next version of the tool allowing the user to generate his own 'questions', generating HTML documents that satisfy those questions.

7. CONCLUSIONS

We have proposed an architecture to produce high quality system-wide information based on source code information. Such documentation will be more effective if code standards are used properly. Our approach uses two different software artifacts to parse and organize the documentation information:

- one is a transformation system, which deals with the information retrieval, and
- the other is a CASE tool that deals with the storage and manipulation of the retrieved information.

Other authors (Wells, Brand and Markosian; 1995; Zoufaly *et al*., 1995; Newcomb and Kotik, 1995) propose the use of transformation systems to help the documentation problem. Our work innovates by using a transformation system together with a repository-based system. For instance, we believe that our proposal, regarding the repository aspect, is related to the work of Edwards and Munro (1993), and Jarzabek and Keam (1995). Jarzabek provides a powerful parser to retrieve information creating a PKB (Program Knowledge Base) and using a wizard generates a domain-knowledge base. Although we did not explore domain-oriented knowledge, having the information in the Talisman repository is a first step to link with other Talisman languages, thus making it possible to handle higher-level representations.

An approach similar to ours was presented by Cross and Hendrix (1995). In their case a source-code grammar was not used, the generation of the documentation was based only on the mark-up grammar alone. Johnson and Erdem (1995) focused attention on the interactive and enquiry aspect of program understanding in an architecture similar to ours.

Future work will not only explore *Documentu's* use, but as mentioned above, explore

the full possible links with other representations dealing with higher levels of abstraction. In particular we are planning to explore the integration with the strategy described by Leite and Cerqueira (1995), which could also be based on the Talisman environment.

Acknowledgements

We would like to thank Marcelo Sant'Anna for everything he has done for the success of this project, since its very beginning.

References

Appelt, W. and Tetteh-Lartey, N. (1993) 'The formal specification of the ISO open document architecture (ODA) standard', *The Computer Journal*, **36**(3), 269–279.

Bigelow, J. (1988) 'Hypertext and CASE', *IEEE Software*, **5**(2), 23–27.

Biggerstaff, T., Mitbarden B. and Webster, D. (1994) 'Program understanding and the concept assignment problem', *Communications of ACM*, **37**(5), 72–83.

Braga C. O., Staa, A. von and Leite, J. C. S. P. (1997) 'A hybrid architecture for documentation production', Technical Report MCC15/97, Departamento de Informática da Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 17 pp.

Brown, H. (1989) 'Standards for structured documents', *The Computer Journal*, **32**(6), 505–514.

Cowan, D. D., Germán, D. M., Lucena, C. J. P. and Staa, A. von (1994) 'Enhancing code for readability and comprehension using SGML', in *Proceedings of the International Conference on Software Maintenance*, IEEE Computer Society Press, Los Alamitos CA, pp. 181–190.

Cross, J. and Hendrix, T. (1995) 'Using generalized markup and SGML for reverse engineering graphical representations of software', in *Proceedings of the Working Conference on Reverse Engineering*, IEEE Computer Society Press, Los Alamitos CA, pp. 2–6.

Edwards, H. and Munro, M. (1993) 'RECAST—reverse engineering from COBOL to SSADM', in *Proceedings of the Working Conference on Reverse Engineering*, IEEE Computer Society Press, Los Alamitos CA, pp. 44–53.

Flanagan, D. (1996) *Java in a Nutshell*, O'Reilly & Associates, Inc., Sebastopol CA, 438 pp.

Garg, P. and Scacchi, W. (1990 'A hypertext to manage software life-cycle documents', *IEEE Software*, **7**(3), 90–98.

Goldfarb, C. (1990) *SGML Handbook*, Oxford University Press Inc., New York NY, 664 pp.

Gosling, J., Joy B. and Steele G. (1996) *Java Language Specification*, Addison-Wesley Publication Co., Reading MA, 825 pp.

Horowitz, E., Kemper, A. and Narasimham, B. (1985) 'A survey of application generators', *IEEE Software*, **2**(1), 40–54.

Ierusalimschy, R., Figueiredo, L. and Celes, W. (1996) 'Lua—an extensible extension language', *Software: Practice and Experience*, **26**(6), 635–652.

Jarzabek, S. and Keam, T. (1995) 'Design of a reverse engineering assistant tool', in *Proceedings of the Working Conference on Reverse Engineering*, IEEE Computer Society Press, Los Alamitos CA, pp. 61–70.

Johnson, W. L. and Erdem, A. (1995) 'Interactive explanation of software systems', in *Proceedings of the 1995 IEEE Knowledge Based Software Engineering Conference*, IEEE Computer Society Press, Los Alamitos CA, pp. 155–164.

Knuth, D. E. (1984) 'Literate programming', *The Computer Journal*, **27**(2), 97–111.

Leite, J. and Cerqueira, P. (1995) 'Recovering business rules from structure analysis specifications', in *Proceedings of the Working Conference in Reverse Engineering*, IEEE Computer Society Press, Los Alamitos CA, pp. 13–21.

Leite, J. C. S. P., Sant'Anna, M. and Freitas, F. (1994) 'Draco-PUC: a technology assembly for domain oriented software development', in *Proceedings of the 3rd IEEE International Conference on Software Reuse*, IEEE Computer Society Press, Los Alamitos CA, pp. 94–101.

Neighbors, J. (1984) 'The Draco approach to constructing software from reusable components', *IEEE Transactions on Software Engineering*, **SE-10**(5), 564–574.

Newcomb, P. and Kotik, G. (1995) 'Reengineering procedural into object-oriented systems', in *Proceedings of the Working Conference on Reverse Engineering*, IEEE Computer Society Press, Los Alamitos CA, pp. 237–252.

Rajlich, V., Damaskinos, N., Khorshid, W. and Linos, P. (1990) 'VIFOR: a tool for software maintenance', *Software: Practice and Experience*, **20**(1), 67–77.

Selfridge, P., Waters, R. and Chikofsky, E. (1993) 'Challenges to the field of reverse engineering', in *Proceedings of the Working Conference in Reverse Engineering*, IEEE Computer Society Press, Los Alamitos CA, pp. 144–150.

Staa, A. von (1993) *Ambiente de Engenharia de Software Talisman, Manual do Usuário*, Staa Informática, Rio de Janeiro, 296 pp.

Staa, A. von, Derraik, A., Braga, C., Costa, G., Kanamori, L., Jaccoud, M., Giovani, P., Hübscher, P., Baptista, R. and Correia, R. (1995) '*Regras e recomendações para a inclusão de especificações no código de programas C ou C++*', Technical Report, Departamento de Informática da Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 25 pp.

Wells, C., Brand, R. and Markosian, L. (1995) 'Customized tools for software quality assurance and reengineering', in *Proceedings of the Second Working Conference on Reverse Engineering*, IEEE Computer Society Press, Los Alamitos CA, pp. 71–77.

World Wide Web Consortium (1995), *Hypertext Markup Language Specification 3.0*, Laboratory of Computer Science, Massachusetts Institute of Technology, Cambridge MA, 170 pp., available via http://www.w3.org

Zoufaly, F., Araya, C., Sanabria, I. and Bendek, F. (1995) 'RESCUE: legacy system translator', in *Proceedings of the Second Working Conference on Reverse Engineering*, IEEE Computer Society Press, Los Alamitos CA, pp. 39–52.

Authors' biographies:

Christiano de Oliveira Braga is a Ph.D. candidate at the Departamento de Informática (Informatics Department) at PUC-Rio. He was the technical manager, designer and main programmer on the *Documentu* project. His research interests are transformation systems, software-engineering environments and formal semantics of programming languages. Christiano received a B.Eng. degree in Computer Engineering from the Pontifical Catholic University of Rio de Janeiro in Brazil (PUC-Rio) in 1993, and an M.S. degree in Computer Science from PUC-Rio in 1996. His email address is: cbraga@inf.puc-rio.br

Arndt von Staa is an Associate Professor at the Departamento de Informática (Informatics Department) at the Pontifical Catholic University of Rio de Janeiro in Brazil (PUC-Rio). He designed and developed the software engineering meta-environment Talisman. His research interests are software quality and process-driven software engineering environments. Currently, he is working on the design and development of a distributed process-driven meta-environment. Arndt got his Ph.D. in Computer Science from the University of Waterloo in Canada. His email address is: arndt@inf.puc-rio.br

Julio Cesar Sampaio do Prado Leite is an Associate Professor at the Departamento de Informática (Informatics Department) at the Pontifical Catholic University of Rio de Janeiro in Brazil (PUC-Rio), and the Director of the Draco-PUC project. His research interests are in the areas of reuse, reverse engineering and requirements engineering, where he performed pioneering work on viewpoint analysis. Julio is a member of the IFIP Working Group 2.9 on software requirements engineering, and has served on various international programming committees, including the IEEE International Conference on Software Reuse and the IEEE International Symposium on Requirements Engineering. He holds a Ph.D. in Computer Science from the University of California, Irvine. His email address is: julio@inf.puc-rio.br